Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding

Year: 2023 | *Citations: 1429* | *Authors: Hang Zhang, Xin Li, Lidong Bing*

Abstract

We present Video-LLaMA a multi-modal framework that empowers Large Language Models (LLMs) with the capability of understanding both visual and auditory content in the video. Video-LLaMA bootstraps cross-modal training from the frozen pre-trained visual and audio encoders and the frozen LLMs. Unlike previous works that complement LLMs to process the visual or audio signals only, Video-LLaMA enables video comprehension by tackling two challenges: (1) capturing the temporal changes in visual scenes, (2) integrating audio-visual signals. To counter the first challenge, we propose a Video Q-former to assemble a pre-trained image encoder into our video encoder and introduce a video-to-text generation task to learn video-language correspondence. For the second challenge, we leverage ImageBind, a universal embedding model aligning multiple modalities, as the pre-trained audio encoder and introduce an Audio Q-former on top of ImageBind to learn reasonable auditory query embeddings for the LLM module. To align the output of both visual and audio encoders with LLM's embedding space, we first train Video-LLaMA on massive video/image-caption pairs and then tune our model with visual-instruction datasets of moderate amount but higher quality. We found Video-LLaMA shows the ability to perceive and comprehend video content and generate meaningful responses grounded in the visual and auditory information presented in the videos.